IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2018, accepted August 6, 2018, date of publication August 17, 2018, date of current version September 21, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2865921

STLoyal: A Spatio-Temporal Loyalty-Based Model for Subway Passenger Flow Prediction

JINZHONG WANG[1,2], XIANGJIE KONG[1], (Senior Member, IEEE), WENHONG ZHAO[3], (Member, IEEE), AMR TOLBA[4,5], ZAFER AL-MAKHADMEH[4], AND FENG XIA[1], (Senior Member, IEEE)

[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China
[2]Shenyang Sport University, Shenyang 110102, China
[3]Ultraprecison Machining Center, Zhejiang University of Technology, Hangzhou 310014, China
[4]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[5]Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Shebin-El-Kom 32511, Egypt

Corresponding authors: Wenhong Zhao (whzhao6666@outlook.com) and Amr Tolba (atolba@ksu.edu.sa)

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group NO (RG-1438-027). This work was also supported in part by the National Natural Science Foundation of China under Grant 61572106, in part by the Natural Science Foundation of Liaoning Province, China, under Grant 201602154, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT18JC09.

ABSTRACT Passenger flow prediction is one of the most important issues in an urban subway system toward smart cities, which can help cut a trip time, plan a trip route, and thus provide a comfortable travel experience. However, there still exist some challenges on how to fully leverage effective and efficient passenger mobility patterns hidden in big traffic data to improve the accuracy of prediction. In this paper, we introduce a concept of the loyal passenger and propose a Spatio-Temporal Loyalty-based model (STLoyal) to improve the precision of prediction through analyzing the characteristics of loyal subway passengers. The proposed STLoyal model is evaluated using real-world subway transaction data sets in Shanghai, China, and it is compared with other state-of-the-art methods. The experimental results show that STLoyal yields superior prediction accuracy on weekdays and weekends in terms of loyalty, time, location, and weather metrics.

INDEX TERMS Passenger loyalty, spatio-temporal analysis, subway passenger flow prediction.

I. INTRODUCTION

Passenger flow prediction is essential in a subway system in smart cities, as it is helpful for planning a trip, thus saving trip time and providing a comfortable travel experience [1]–[3]. With the rapid development of mobile computing and sensing technologies, the emergence of big traffic data provides a unique opportunity for developing effective and efficient algorithms and extracting latent mobility patterns to improve the accuracy of flow prediction.

Emanating from the large scale and heterogeneous traffic data, passenger flow prediction receives wide attentions from the research community. Existing passenger flow prediction approaches can be categorized into three types, i.e. time-series approaches (autoregressive integrated moving average model) [4], [5], non-parametric approaches (deep learning) [6]–[9], and hybrid methods (combination of the above-mentioned two types) [10], [11]. However, hand-engineered features may be error-prone, which impact the prediction results. Thus, great challenges still exist in extracting important mobility patterns to improve prediction accuracy. Passenger flow prediction is a subtopic of traffic flow prediction, and it plays a significant role in a variety of applications such as smart cities [12]–[14], intelligent transportation services [15]–[18], trip planning [19]–[21], and vehicular social networks [22]–[24].

To the best of our knowledge, existing research focuses on the improvement of prediction accuracy by analyzing human mobility patterns [25]–[28]. Reference [16] exploits the relationship between boarding passengers and departing passengers in combination with two additional features to improve the performance of passenger flow prediction. Reference [29] focuses on how to identify pickpocket from large-scale transit records by analyzing historical travel behaviors, social comparisons, and current travel behaviors. Reference [30] utilizes accelerometers on smartphones to track the trajectories of passengers in a subway network and obtained 89% accuracy for trips associated with four stations. Reference [31] aggregates passengers that have similar boarding times into a cluster and identifies urban human mobility patterns such as sporadic usage and commute practices.

The above-mentioned studies share insights of mobility patterns of urban transit users and provide a valuable reference for analyzing passenger travel behaviors. However, these studies do not consider the loyalty of transit users. The loyalty of subway passengers mainly reflects the regularities of travelling by subway in their daily life and is also a characteristic of human mobility patterns in urban transit. Loyal passengers refer to the persons who often get on and off the same stations at the same time slots. In this paper, we analyze the phenomena and define it as loyalty. From a sociological point of view, passenger loyalty reflects the tie strength between passengers and stations to some extent. Loyal passengers are more likely to take the subway in the immediate future. This information is very helpful to predict future ridership. Thus, a study of passenger loyalty should be further investigated.

Some scholars mainly focus on how to utilize behavior characteristics of passengers to improve the accuracy of passenger flow prediction. Reference [32] proposes personality traits related to travel behaviors in public transport and verifies the importance of habit and intention for the predictability. Reference [10] pays close attention when predicting passenger flow under mega events and develops a hybrid algorithm including hashtag based event detection and convex optimization based prediction. References [33]–[35] utilize spatio-temporal knowledge to predict crowd flow on diverse spatio-temporal datasets.

In contrast to the existing research, we propose a data-driven Spatio-Temporal Loyalty-based model (STLoyal) to solve the issue above. We are the first to introduce the concept of loyal passengers. We subsequently detect the number of loyal passengers in each station, and acquire travel characteristics of passengers by subway (loyalty, time, location, and weather). Subsequently, the four features are fed into the STLoyal model to predict passenger flow at every time slot on weekdays and weekends, respectively.

A definition of 'loyal passenger' is proposed to verify whether or not strong ties exist between the volume of loyal passengers and the number of all passengers in an urban subway network. STLoyal model is then proposed including Loyal Passenger Detection algorithm (LoPaD) and Multiple Factors Combined Prediction algorithm (MFCP), which can identify the four-dimensional properties of passenger travel behaviors which are used to forecast the volume of boarding passengers for subway stations at different time slots. Furthermore, we compare STLoyal over Support Vector Regression (SVR), Back-Propagation Neural Network (BPNN), and Gradient Boost Decision Tree (GBDT) in terms of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Normalized Mean Absolute Error (NMAE), to verify the performance of STLoyal.

II. METHODS

A. OVERVIEW OF STLoyal

As shown in Fig.1, we firstly preprocess the datasets and clean erroneous information generated by transmission delay and device fault. Subsequently, we analyze the subway mobility patterns of Shanghai, extract the spatio-temporal and loyalty characteristics during different time slots, and sense the distribution of loyal passengers for each subway station. Meanwhile, we propose a data-driven STLoyal model to predict subway passenger flow. STLoyal is a two-step predictive model which consists of LoPaD and MFCP.

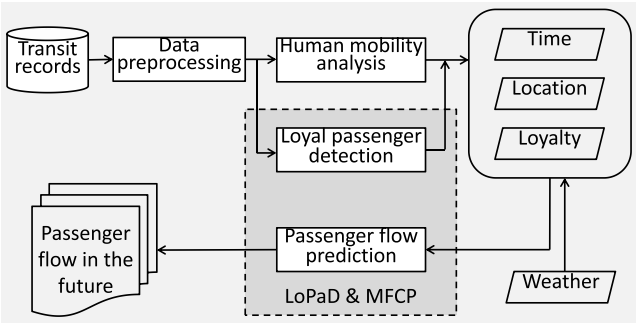

FIGURE 1. Framework of STLoyal. The light grey area shows all of the modules of STLoyal. Specifically, we propose a definition of loyal passenger based on the patterns of passenger mobility. We utilize the four metrics (loyalty, time, location, and weather) to improve accuracy of prediction.

B. LOYAL PASSENGER DETECTION (LOPAD)

Loyalty illustrates human behavior from an individualistic aspect. Intuitively, loyalty is an economic indicator and plays an important role in several applications. To the best of our knowledge, loyalty is originally measured using three aspects: satisfaction, repeat purchase, and probability of recommending it to others [36]. In this paper, we propose the metric (passenger loyalty) to improve the accuracy of predicting subway passenger flow in the future.

In public transit services, subway transaction records include passengers' card ID, boarding time and station, departing time and station, and trip fare. This information reflect human mobility patterns from spatial and temporal dimensions. In other words, we are able to obtain the frequency of the use of public transit services, which is closely associated with loyalty of passengers. Although complaints exist with regard to the volume of passenger flow and a longer time for transfer during peak hours, some people prefer to travel by subway in terms of cost and efficiency. The high frequency of subway use reflects a strong passenger loyalty.

In practice, people usually follow their work routine and therefore embark and disembark the same stations with some degree of certainty. From the perspective of a subway station, the frequency of interaction of loyal passengers is higher than irregular passengers. We define the riders with the above-mentioned characteristics as loyal passengers. Based on empirical evidence, we propose a detection algorithm (LoPaD) to distinguish loyal passengers from transaction records. Let j denote the jth time slot and i denotes a station, we define passenger loyalty threshold λ_{ij} in (1), which is potentially helpful to represent subway human mobility patterns and is utilized to identify loyal passengers.

Algorithm 1 Pseudocode of LoPaD

Input: P_{ij}
Output: L_{ij}

procedure :generate the number of loyal passengers for each station
  $P_{ij} \leftarrow 0$
  $Sum \leftarrow 0$
  $L_{ij} \leftarrow 0$
  for each station $i \in S$ **do**
    for each time slot $j \in T$ **do**
      for $k \in [D_0, D_1, ..., D_n]$ **do**
        $Sum = Sum + k \times P_{ijk}$
        $P_{ij} = P_{ij} + P_{ijk}$
      end for
      $\langle H_{ij} \rangle = Sum/P_{ij}$
      $\lambda_{ij} = (\langle H_{ij} \rangle - 1) \times \ln(P_{ij})$
      if $k \geq \lambda_{ij}$ **then**
        $L_{ij} = L_{ij} + P_{ijk}$
      end if
    end for
  end for
  return L_{ij}
end procedure

Definition 1 (Loyal Passenger): For a station i, a passenger p is loyal if and only if: $p.origin = S_i, p.v_{ij} \geq \lambda_{ij}$, and $MER(P) \leq d$, where λ_{ij} and v_{ij} are the loyalty threshold and the times of visit at the jth time slot, respectively. $MER(p)$ denotes the temporal minimum enclosing rectangle of p. d is the corresponding observation period.

$$\lambda_{ij} = (\langle H_{ij} \rangle - 1) \ln(P_{ij}), \tag{1}$$

$$\langle H_{ij} \rangle = \frac{\sum_{k=D_0}^{D_n} k \times P_{ijk}}{P_{ij}}, \tag{2}$$

where $\langle H_{ij} \rangle$ and P_{ij} denote the average frequency of visits and the number of passengers visiting k times in the jth time slot for the ith station. D_0 and D_n represent the starting day and ending day of the observation period respectively.

Detection has the advantage that λ_{ij} is nonparametric, depending only on the observables $\langle H_{ij} \rangle$ and P_{ij}. Thus, the passenger loyalty threshold is related to $\langle H_{ij} \rangle - 1$ (-1 arises as the minimum H_{ij} value is set to 1). Meanwhile, we introduce a logarithmic factor in P_{ij} to reflect sample size dependence. The proposed LoPaD is illustrated in Algorithm 1.

C. MULTIPLE FEATURES COMBINED PREDICTION (MFCP)

1) TIME

As a temporal feature, time is closely related to the variation of passenger flow and has a direct impact on prediction accuracy. We conduct extensive experiments on weekdays and weekends at different time slots, respectively. To be specific, we set τ minute(s) as an interval and the jth time slot is defined according to (3):

$$t_j = [j\tau, (j+1)\tau), \tag{3}$$

where $j = 1, 2, \ldots, (120/\tau) - 1$. Based on the subway transaction dataset, we define the time interval to be 10 minutes and categorize transaction records into 120 time slots form 4:00 to 23:00.

2) WEATHER

Weather is an important factor in impacting the change of passenger flow and may include heavy snow, torrential rain, dense fog. In terms of safety and travel time, inclement weather conditions may encourage public transit instead of private cars thereby resulting in a larger passenger flow than when weather conditions are fair. Weather data consists of temperature (tem), visibility (vis), wind power (win), and general conditions (con) as shown in (4). All of these items are normalized to real numbers in the range of [0, 1]. Specifically, we utilize $W_{D_k}^{S_i}$ to represent the weather conditions in the ith subway station during the kth day and illustrate it as follows:

$$W_{D_k}^{S_i} = (W_{t_0}^{i,k}, W_{t_1}^{i,k}, \cdots, W_{t_j}^{i,k})^T, \tag{4}$$

where $W_{t_j}^{i,k}$ is the weather vector at the ith station from jth time slot to $(j+1)$th time slot on the kth day and is represented in (5).

$$W_{t_j}^{i,k} = (tem_{i,j}, vis_{i,j}, win_{i,j}, con_{i,j})^T. \tag{5}$$

As shown in (6), the matrix W, of n columns and m rows, is used to represent the weather conditions at different stations during different periods. m and n denote the number of days and stations respectively. Specifically, m ranges from 1 to 30 and n is from 1 to 288.

$$\begin{aligned} W &= (W_1, W_2, \cdots, W_n) \\ &= \begin{pmatrix} W_{D_1}^{S_1} & W_{D_1}^{S_2} & \cdots & W_{D_1}^{S_n} \\ W_{D_2}^{S_1} & W_{D_2}^{S_2} & \cdots & W_{D_2}^{S_n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{D_m}^{S_1} & W_{D_m}^{S_2} & \cdots & W_{D_m}^{S_n} \end{pmatrix} \end{aligned} \tag{6}$$

3) LOCATION

As a spatial index, location also plays an important role in passenger flow prediction. In other words, subway stations have their own characteristics and functions due to their location and POI configurations. It should be noted that POI is an abbreviation for point of interest and denotes a specific location that someone may be interested in, such as subway stations, hotels, bars, supermarkets, or any other categories used in urban cities. In this paper, we formulate POI as a interest point around subway stations.

We utilize $G_{D_k}^{S_i}$ to represent the number of boarding passengers in the ith subway station on the kth day and illustrate it as follows:

$$G_{D_k}^{S_i} = (G_{t_0}^{i,k}, G_{t_1}^{i,k}, \cdots, G_{t_j}^{i,k})^T, \tag{7}$$

where $G_{t_j}^{i,k}$ is the number of boarding passengers at the ith station from jth time slot to $(j+1)$th time slot of the kth day.

As shown in (8), the matrix G, of n columns and m rows, is used to represent the distribution of on-boarding passengers at different stations during various periods.

$$G = (G_1, G_2, \cdots, G_n) = \begin{pmatrix} G_{D_1}^{S_1} & G_{D_1}^{S_2} & \cdots & G_{D_1}^{S_n} \\ G_{D_2}^{S_1} & G_{D_2}^{S_2} & \cdots & G_{D_2}^{S_n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{D_m}^{S_1} & G_{D_m}^{S_2} & \cdots & G_{D_m}^{S_n} \end{pmatrix} \tag{8}$$

4) LOYALTY

Loyalty is a novel and important factor when improving the accuracy of predicting passenger flow from the perspective of passenger behaviors. According to the definition of loyalty outlined above, loyal passengers often travel more regularly than ordinary passengers in terms of their daily trajectories. Therefore, we introduce a matrix L to represent the number of loyal passengers in all subway stations during different time slots and L_{ij} denotes the number of loyal passengers at the ith station on the jth time slot.

$$L = (L_1, L_2, \cdots, L_n) = \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{m1} & L_{m2} & \cdots & L_{mn} \end{pmatrix} \tag{9}$$

Then, we propose the Multiple Factors Combined Prediction (MFCP) algorithm to predict the volume of passengers by utilizing multiple linear regression and multiple travel features. As a low computational cost machine learning algorithm, MFCP attempts to exploit the relationship between four explanatory variables and a dependent variable by finding a suitable equation for observed data. The general form of the MFCP method is:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p + \varepsilon, \tag{10}$$

where $\alpha = [\alpha_0, \alpha_1, \cdots, \alpha_p]^T$ denotes the metric regression coefficient matrix. ε refers to an independent identically distributed residual error. Y is the expected value represented as a function of p-number of independent variables. In addition, we transform (10) to a vector form as shown in the following formula.

$$Y = \alpha X + \varepsilon, \tag{11}$$

where α is the vector of the coefficients and is estimated by the least squares method which minimizes the square sum of random errors ε. The least-square estimator of the regression coefficient vector is:

$$\hat{\alpha} = (X^T X)^{-1} X^T y = (\sum x_i x_i^T)^{-1} (\sum x_i y_i), \tag{12}$$

where y_i denotes the ith value of the dependent variable, x_i represents the ith value of the explanatory variables.

In the experiment, we firstly acquire the four features (loyalty, time, location, and weather) of subway stations, which are denoted by the corresponding matrix. Then we feed these features into our proposed MFCP. Finally, we are able to predict the volume of boarding passengers with suitable factors.

III. EXPERIMENTS

A. DATASET PREPROCESSING

Subway transaction data is obtained from Shanghai. The data collected spans over the whole month of April, 2015. As shown in Table 1, the dataset consists of eight fields such as card number, swipe time, fare, and geographical coordinates. For the statistics part, the dataset contains 288 subway stations, 47 of which are transfer stations.

TABLE 1. Description and statistics of subway card transaction data.

Name	Field	Annotation	Example
Transaction data	CardID	Smart card ID	2201252167
	Date	Swipe card date	2015-04-01
	Time	Swipe card time	08:19:00
	Line	Subway line ID	7
	Station	Station name	Shangda Road
	Fee	Subway fare	4
	Longitude	Coordinates	121.465545
	Latitude	Coordinates	31.224068
Statistics	Time	April, 2015	1st-30th
	Days	30	21 weekdays
	Line	14	No.1-13&16
	Station	288	47 transfers

The raw dataset collected from the automatic fare collection system contains a total of more than 451 million trading records of 14 subway lines. First, we preprocessed the dataset by deleting repeated and dirty data of abnormal values (about 0.7%). Second, we extracted trip information of every passenger, i.e. origin, destination, distance, duration, and interval by utilizing data manipulation language. Considering the characteristics of human mobility, we separated weekdays and weekends in the dataset for analysis.

From the perspective of an individual, passengers usually travel from one station to another at different time slots thereby highlighting latent spatio-temporal features. The boarding volume (particularly that of loyal passengers) of each station is extracted for each time interval.

B. IDENTIFYING LOYAL PASSENGERS

We utilize the LoPaD algorithm which is part of our proposed STLoyal model to identify the number of loyal passengers for each subway station at different time intervals. There are 288 subway stations in our dataset. Based on the POI data, we analyze the characteristics of functional regions around stations and divide them into eight kinds of functional clusters [37], i.e. railway stations, commercial areas, entertainment areas, residential regions, scene spots, and areas of

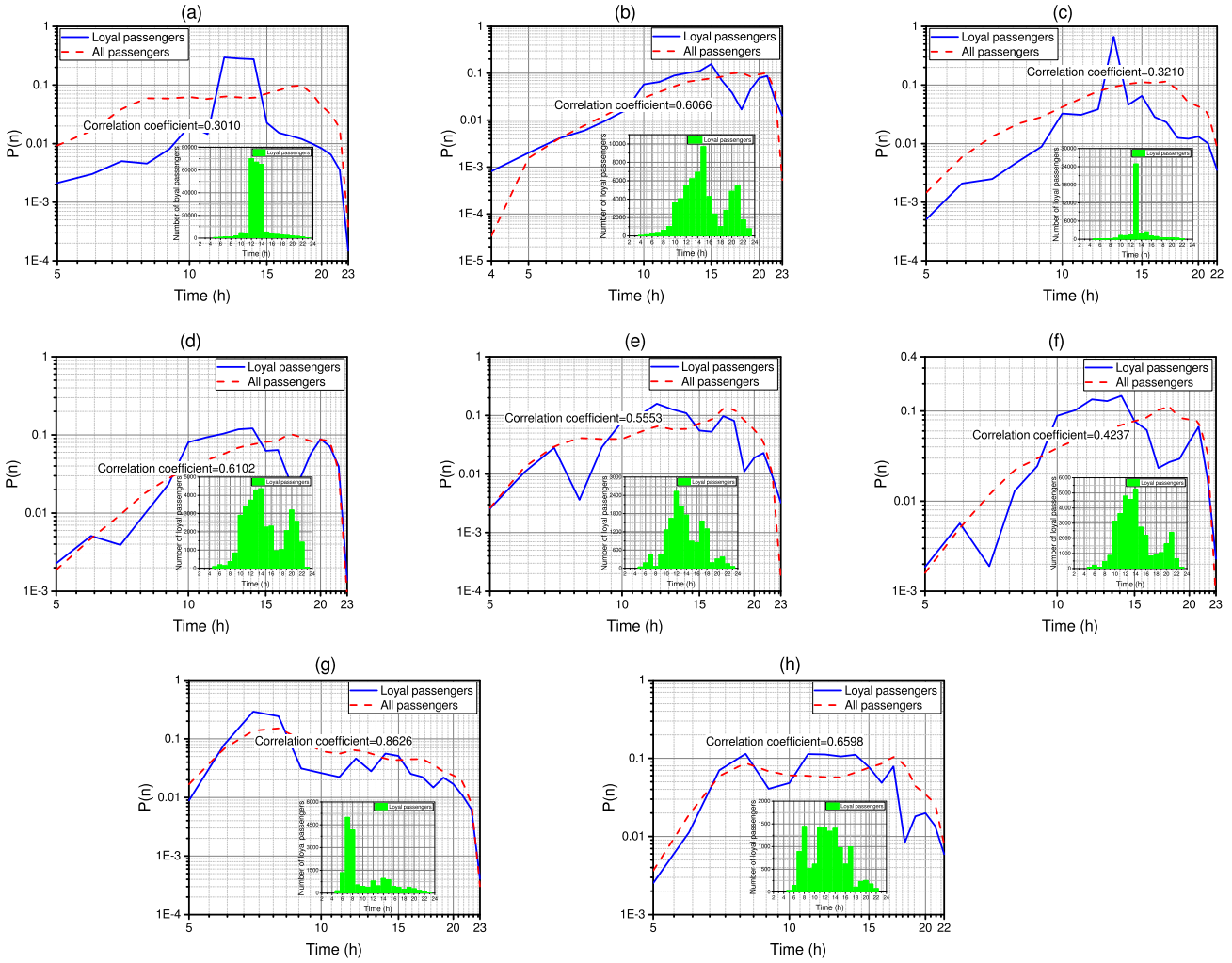

FIGURE 2. **The probability distribution of loyal passengers and all passengers in log-log coordination. The red and blue lines represent the probability distribution of loyal passengers and all passengers from 5:00 to 23:00, respectively. The green bar charts indicate the number of loyal passengers over time. (a) Shanghai Railway Station. (b) People Square Station. (c) Yu Garden Station. (d) East Nanjing Road Station. (e) Zhangjiang High-Tech Park. (f) Jing'an Temple Station. (g) Tonghe Newly State Station. (h) Central Yanggao Road Station.**

historic interest. For a clearer understanding, we select eight representative subway stations to explore the distribution of loyal passengers, namely; Shanghai Railway Station, People Square Station, Yu Garden Station, East Nanjing Road Station, Zhangjiang High-Technology Park, Jing'an Temple Station, Tonghe Newly State Station, and Central Yanggao Road Station, as shown in Figs. 2(a) - 2(h). More importantly, we verify whether or not a strong correlation exists between the number of loyal passengers and the volume of all riders.

The distribution of passenger flow over time is shown in Fig. 2 with a histogram subgraph to illustrate the composition of loyal passengers. We notice that Tonghe Newly State Station which is located in residential regions has a peak value of loyal passengers at 7:00 am. in Fig. 2(h), whereas the other seven stations climb to a peak value at 12:00-14:00 in the afternoon as shown in Figs. 2(a) - 2(g). This reflects the difference on spatio-temporal mobility patterns as it relates to service functions. In addition, the volume of loyal passengers is also stationary and very useful for future passenger flow prediction.

To measure the relationship between loyal passengers and all passengers, we utilize a correlation coefficient to quantify the type of dependence using fundamental statistics. A correlation coefficient is a statistical indicator, evaluating the relationship between two variables or two observations. 1 denotes the strongest relationship between two variables, while 0 denotes no linear relationship between two variables.

Through statistical analysis, we discover that the number of loyal passengers shows a strong correlation with the volume of all passengers in five stations as shown in Figs. 2(b), 2(d), 2(e), 2(g), and 2(h). The other three stations represent a moderate correlation with a coefficient value between 0.3 and 0.5 referred to Figs. 2(a), 2(c), and 2(f). Based on the above-mentioned description, we conclude that the number of loyal passengers has a close relationship to the number of all passengers. In other words, loyalty is a valuable

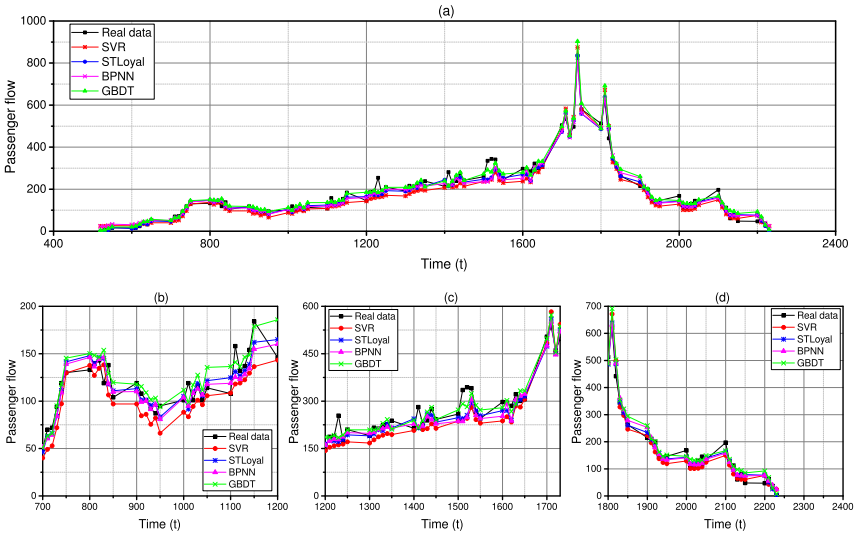

FIGURE 3. The predictive results of four models for Yuyuan Station on a weekday. (a) The overall predictive results on Apr. 30th. (b) The predictive results from 7:00-12:00. (c) The predictive results from 12:00 to 17:00. (d) The predictive results from 18:00 to 24:00.

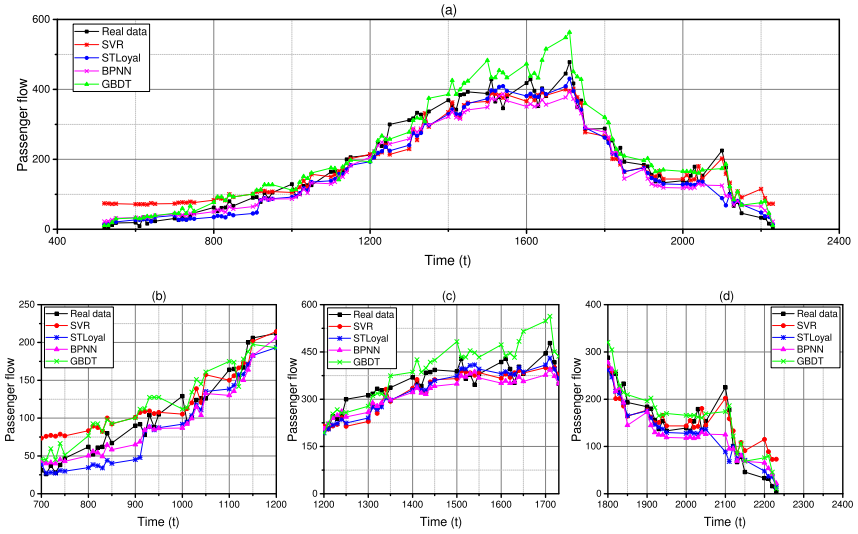

FIGURE 4. The predictive results of four models for Yuyuan Station on a weekend. (a) The overall predictive results on Apr. 26th. (b) The predictive results from 7:00-12:00. (c) The predictive results from 12:00 to 17:00. (d) The predictive results from 18:00 to 24:00.

characteristic when improving the accuracy of passenger flow prediction.

C. PREDICTING PASSENGER FLOW

Based on a series of experiments, we compare the performance of our proposed STLoyal model with SVR, BPNN, and GBDT. Specifically, SVR model mainly utilizes a kernel function to represent the training samples as points in space. BPNN model utilizes back propagation algorithm to predict future volume of boarding passengers; while GBDT model improves prediction accuracy by using a gradient boosting method.

Based on our acquired multiple features, we utilize our proposed STLoyal model to conduct extensive experiments. A different training set is utilized to forecast subway passenger flow on weekdays and weekends. To be specific, we utilize the first 16 weekdays data as a training set and the last 5 weekdays as a test set, whereas we take the first 6 weekends as a training set and the last 3 weekends as a test set. We then use 10 minutes as a time slot to sense the regularity of passenger flow during different periods.

According to the empirical analysis, we show the predictive results of two subway stations, i.e. Yuyuan station and Tonghe station. As shown in Fig. 3 and Fig. 4, we discover the characteristics of passenger distribution on weekdays and weekends. The passenger volume for Yuyuan Station has a greater fluctuation on weekdays than on weekends

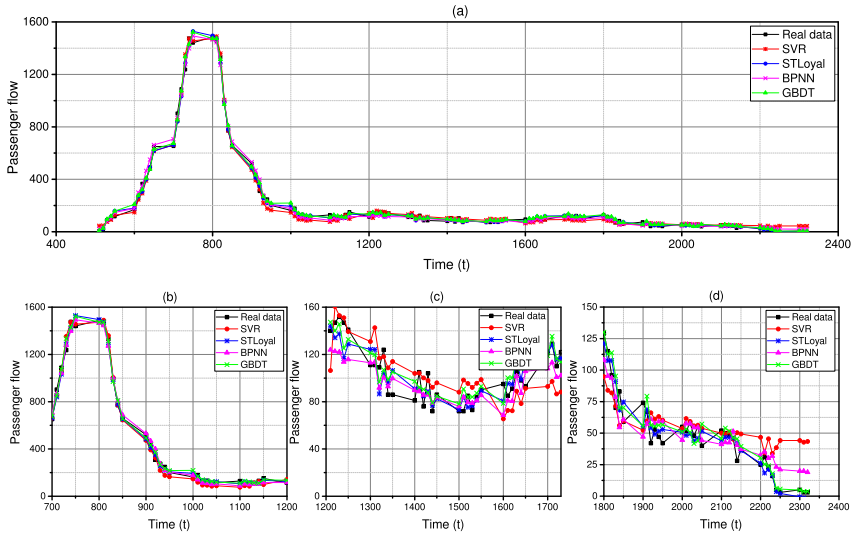

FIGURE 5. The predictive results of four models for Tonghe Station on a weekday. (a) The overall predictive results on Apr. 30th. Its specific predictive results in three selected time periods (b), (c), and (d).

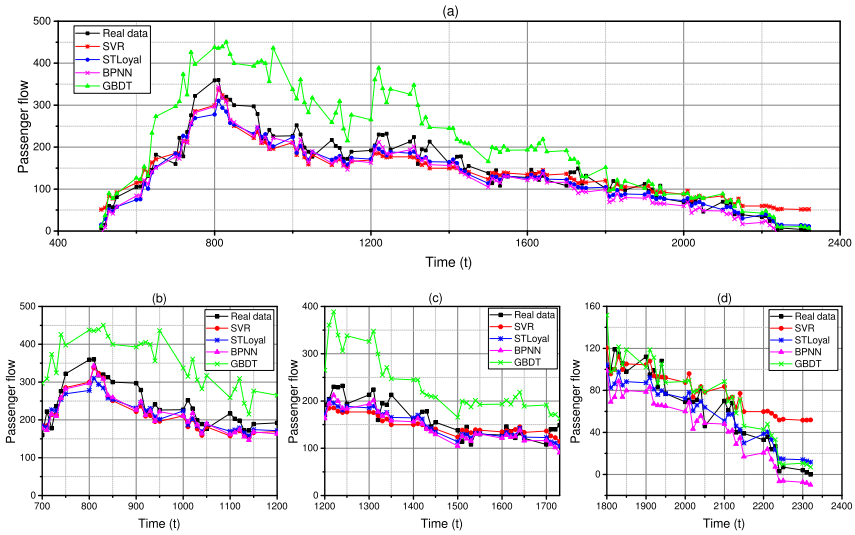

FIGURE 6. The predictive results of four models for Tonghe Station on a weekend. (a) The overall predictive results on Apr. 26th. Its specific predictive results in three selected time periods (b), (c), and (d).

particularly during evening rush hours. Meanwhile, the predictive precision of four models on weekdays are superior to that on weekends, which are related to various social activities for people by the end of this week.

As shown in Figs. 3(b) - 3(d) and Figs. 4(b) - 4(d), the blue curve represents the performance of STLoyal in different time slots. In Fig. 4(b), STLoyal has a lower prediction accuracy at 7:00 - 9:00, whereas it demonstrates the best performance during the other periods. In general, SVR has a larger prediction error than the other three models.

As shown in Fig. 5 and Fig. 6, passenger flow reaches a peak at 8:00 and goes down instantly, which is associated with the station's service function in residential regions. In addition, the volume of passenger flow on weekdays is four times that on weekends. However, for relatively regular mobility patterns, the prediction models have a better performance on weekdays than on weekends. For a clearer illustration of predictive results, the graphs are enlarged (zoomed-in) in Figs. 5(b) - 5(d) and Figs. 6(b) - 6(d). Through horizontal and vertical comparative analysis, we conclude STLoyal is superior to the other 3 models in terms of accuracy and stability.

We also list prediction results of four algorithms for subway passenger flow as shown Table 2. STLoyal achieves the prediction accuracy of 90.92% at weekdays and 81.6% at weekends respectively. Based on passenger loyalty and

TABLE 2. Prediction accuracy of four models.

Time	SVR	STLoyal	BPNN	GBDT
Weekdays	37.36%	**90.92%**	57.10%	43.31%
Weekends	21.93%	**81.64%**	39.48%	18.82%

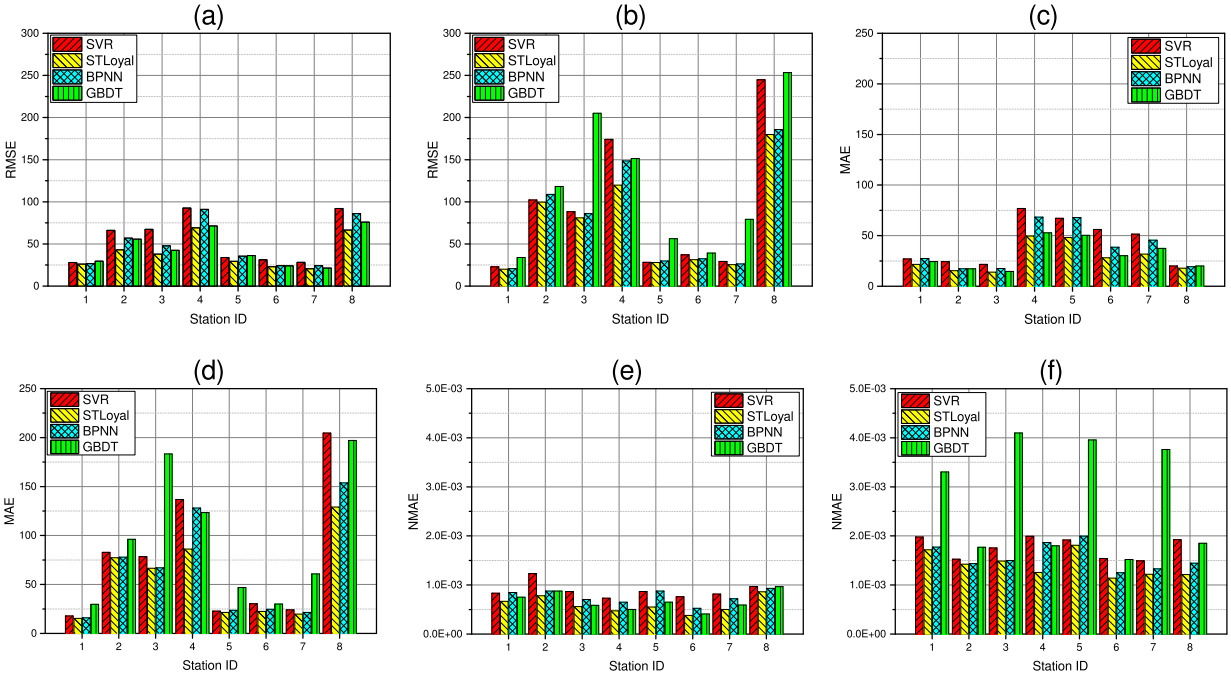

FIGURE 7. RMSE, MAE, and NMAE of four models on weekdays and weekends. X-axis represents the number of stations with various functions. (a) RMSE on weekdays. (b) RMSE on weekends. (c) MAE on weekdays. (d) MAE on weekends. (e) NMAE on weekdays. (f) NMAE on weekends.

multiple features combined prediction, STLoyal outperforms other state-of-the-art methods.

D. PERFORMANCE EVALUATION

1) EVALUATION METRICS

To compare the performance of our proposed STLoyal model with other methods, we utilize the following three indices RMSE, MAE, and NMAE.

- Root Mean Square Error (RMSE). RMSE is defined as the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |f_i - y_i|^2}{n}} \tag{13}$$

  where f_i and y_i denote predicted and real values respectively, and n is defined as the number of measurements.
- Mean Absolute Error (MAE). In this paper, MAE is given by:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| \tag{14}$$

- Normalized Mean Absolute Error (NMAE). This metric is the normalized form of MAE and is represented as follows:

$$NMAE = \frac{MAE}{\sum_{i=1}^{n} y_i} \tag{15}$$

2) COMPARISON OF RESULTS

We demonstrate three performance indices of all four predictive models in eight representative stations of different service functions on weekdays and weekends, respectively. As shown in Fig. 7, we discover that the performance of four models on weekdays outperforms that on weekends, as they are dependent on people regular mobility patterns.

As shown in Fig. 7(a) and Fig. 7(b), STLoyal has the lowest RMSE in all four models for eight representative stations. Moreover, referring to Table 3, RMSE in STLoyal model are 33.74 on weekdays and 73.15 on weekends accordingly. For the other three comparative models, RMSE in GBDT is 40.62 on weekdays and RMSE in BPNN is 82.77 on weekends. Vertical comparative analysis demonstrates that STLoyal has a high prediction accuracy as a result of introducing passenger loyalty.

As seen from Fig. 7(c) and Fig. 7(d), the MAE of STLoyal is the lowest among all four prediction models both on

TABLE 3. Performance of four algorithms with loyalty.

Model	RMSE		MAE		NMAE	
	Weekdays	Weekends	Weekdays	Weekends	Weekdays	Weekends
SVR	54.97	90.83	43.07	74.69	8.84E-4	1.76E-3
STLoyal	**33.74**	**73.15**	**23.54**	**63.00**	**4.81E-4**	**1.41E-3**
BPNN	49.04	82.77	37.73	67.06	7.66E-4	1.57E-3
GBDT	40.62	137.06	26.88	115.89	5.67E-4	3.16E-3

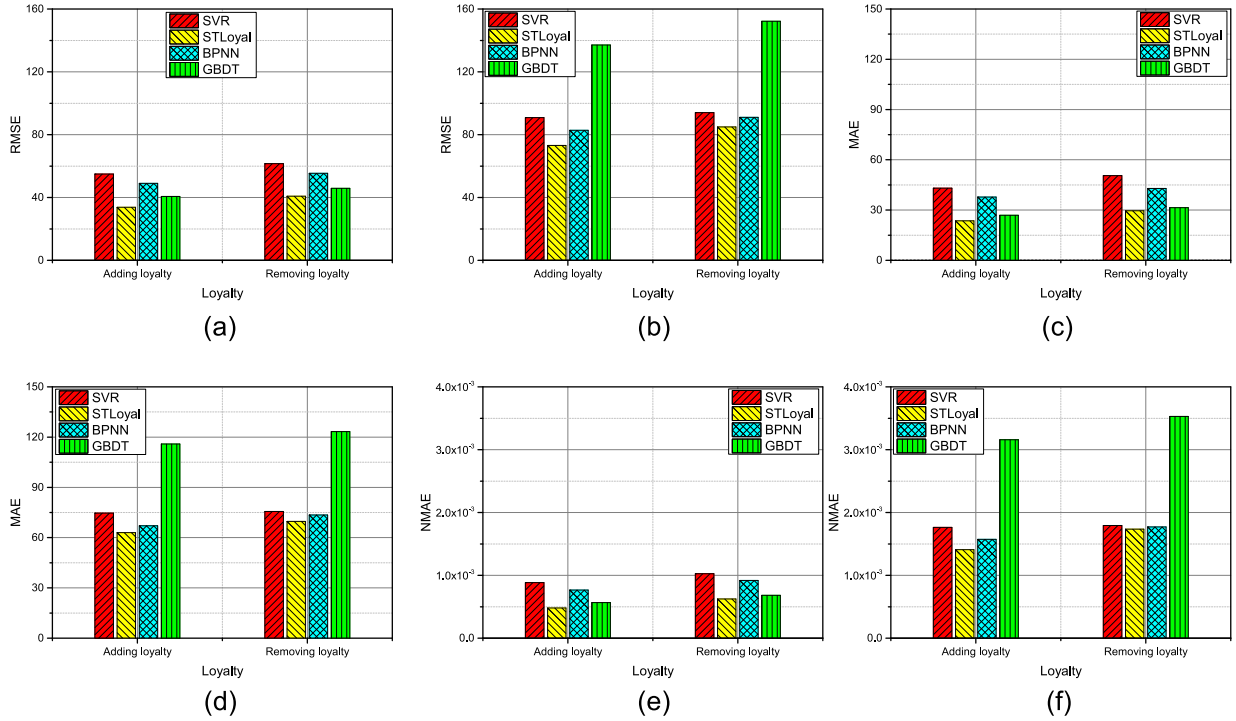

FIGURE 8. Loyalty contribution analysis on weekdays and weekends. The figure illustrates the effects of omitting and including the loyalty index for four predictive models, respectively. (a) RMSE on weekdays. (b) RMSE on weekends. (c) MAE on weekdays. (d) MAE on weekends. (e) NMAE on weekdays. (f) NMAE on weekends.

TABLE 4. Statistical analysis of loyalty contribution (decreasing error rate).

Model	RMSE		MAE		NMAE	
	Weekdays	Weekends	Weekdays	Weekends	Weekdays	Weekends
SVR	10.59%	3.38%	14.83%	1.23%	14.16%	1.68%
STLoyal	**17.45%**	**13.81%**	**20.51%**	**9.61%**	**22.87%**	**18.97%**
BPNN	11.47%	9.08%	12.11%	8.75%	16.66%	11.30%
GBDT	11.38%	9.95%	14.48%	5.98%	16.85%	19.48%

weekdays and weekends, which demonstrates a good performance once again. To be specific, the MAE of STLoyal is 23.54 on weekdays and 63.00 on weekends as indicated in Table 3. Furthermore, our statistical results show that urban human activities on weekends are more complicated than that on weekdays.

As shown in Fig. 7(e) and Fig. 7(f), the value of NMAE generally maintains a slight variation in addition to the GBDT model on weekends, which is correlated to the normalized operation. The NMAE of STLoyal has the lowest value with 0.48% on weekdays and 1.41% on weekends in Table 3. GBDT model ranks second with 5.67% on weekdays, whereas BPNN is runner-up with 1.57% on weekends.

To verify the effectiveness of the loyalty metric, we also analyze the predictive accuracy by adding and removing loyalty on weekdays and weekends, respectively. As shown in Fig. 8, we can discover the positive effect of loyalty index for passenger flow prediction. In Figs. 8(a) and 8(b), the

3.38 - 23.09 percent drop in RMSE demonstrated by the addition of loyalty indicates that the metric plays an important role in predicting passenger flow. From Figs. 8(c) and 8(d), we see that loyalty is critically important for improving predictive accuracy. Figs. 8(e) and 8(f) show that NMAE values also decrease for all predictive models by the addition of loyalty. Table 4 lists the statistical results for loyalty contribution on weekdays and weekends, where each data represents the percentage of reduced predictive error.

In summary, by utilizing the proposed STLoyal, we discover that the experimental results have higher prediction accuracy than that of the other three models in eight representative stations.

IV. CONCLUSION

In this paper, we have developed a Spatio-Temporal Loyalty-based predictive model named STLoyal by utilizing the loyalty of subway passengers and three other travel features (time, location, and weather). We used a real-world transaction dataset to compare STLoyal with three models (SVR, BPNN, and GBDT). Our results indicate the best performance of STLoyal on weekdays and weekends. We have conducted extensive experiments on analyzing the contribution of passenger loyalty, which decreases the predictive error by 3.38% - 17.45%. STLoyal possesses the lowest prediction error rate with a NMAE of 0.48% on weekdays and 1.41% on weekends. We also utilized the evaluation indications such as RMSE, MAE, and NMAE, to verify the stability and effectiveness of STLoyal.

REFERENCES

[1] C. Martani, S. Stent, S. Acikgoz, K. Soga, D. Bain, and Y. Jin, "Pedestrian monitoring techniques for crowd-flow prediction," *Proc. Inst. Civil Eng.-Smart Infrastruct. Construct.*, vol. 170, no. 2, pp. 17–27, 2017

[2] T. Liebig, N. Piatkowski, C. Bockermann, and K. Morik, "Predictive trip planning—Smart routing in smart citie," in *Proc. EDBT/ICDT Workshops*, 2014, pp. 331–338.

[3] F. Xia, A. Rahim, X. Kong, M. Wang, Y. Cai, and J. Wang, "Modeling and analysis of large-scale urban mobility for green transportation," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1469–1481, Apr. 2018.

[4] C. Liu, S. C. H. Hoi, P. Zhao, and J. Sun, "Online ARIMA algorithms for time series prediction," in *Proc. AAAI*, 2016, pp. 1867–1873.

[5] E. Ko, J. Ahn, and E. Y. Kim, "3D Markov process for traffic flow prediction in real-time," *Sensors*, vol. 16, no. 2, p. 147, 2016.

[6] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018.

[7] L. Liu and R.-C. Chen, "A novel passenger flow prediction model using deep learning methods," *Transp. Res. C, Emerg. Technol.*, vol. 84, pp. 74–91, Nov. 2017.

[8] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI*, 2017, pp. 1655–1661.

[9] X. Song, H. Kanasugi, and R. Shibasaki, "Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level," in *Proc. IJCAI*, 2016, pp. 2618–2624.

[10] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1623–1632, Jun. 2017.

[11] Q. Shang, C. Lin, Z. Yang, Q. Bing, and X. Zhou, "A hybrid short-term traffic flow prediction model based on singular spectrum analysis and kernel extreme learning machine," *PLoS ONE*, vol. 11, no. 8, p. e0161259, 2016.

[12] B. Tang *et al.*, "Incorporating intelligence in fog computing for big data analysis in smart cities," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2140–2150, Oct. 2017.

[13] J. Wang, X. Kong, A. Rahim, F. Xia, A. Tolba, and Z. Al-Makhadmeh, "Is2fun: Identification of subway station functions using massive urban data," *IEEE Access*, vol. 5, pp. 27103–27113, 2017.

[14] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa, "Visual exploration of changes in passenger flows and tweets on mega-city metro network," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 85–99, Mar. 2016.

[15] S. Li, L. Yang, and Z. Gao, "Optimal switched control design for automatic train regulation of metro lines with time-varying passengers arrival flow," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 425–440, Jan. 2018.

[16] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[17] D. Chen, "Research on traffic flow prediction in the big data environment based on the improved RBF neural network," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2000–2008, Aug. 2017.

[18] X. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, "LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data," *World Wide Web*, vol. 21, no. 3, pp. 825–847, 2017.

[19] A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman, "Public transport trip purpose inference using smart card fare data," *Transp. Res. C, Emerg. Technol.*, vol. 87, pp. 123–137, Feb. 2018.

[20] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Generat. Comput. Syst.*, vol. 61, pp. 97–107, Aug. 2016.

[21] W. Gu, M. Jin, Z. Zhou, C. J. Spanos, and L. Zhang, "MetroEye: Towards fine-grained passenger tracking underground," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, 2016, pp. 77–80.

[22] Z. Ning, F. Xia, N. Ullah, X. J. Kong, and X. P. Hu, "Vehicular social networks: Enabling smart mobility," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 16–55, May 2017.

[23] T. Xu *et al.*, "Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1285–1294.

[24] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, and C. Liu, "Exploring human mobility patterns in urban scenarios: A trajectory data perspective," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 142–149, Mar. 2018.

[25] J. Zhao, C. Tian, F. Zhang, C. Xu, and S. Feng, "Understanding temporal and spatial travel patterns of individual passengers by mining smart card data," in *Proc. 17th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 2991–2997.

[26] L. Sun and J. G. Jin, "Modeling temporal flow assignment in metro networks using smart card data," in *Proc. 18th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 836–841.

[27] F. Zhang, J. Zhao, C. Tian, C. Xu, X. Liu, and L. Rao, "Spatiotemporal segmentation of metro trips using smart card data," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1137–1149, Mar. 2016.

[28] X. Zhan, Y. Zheng, X. Yi, and S. V. Ukkusuri, "Citywide traffic volume estimation using trajectory data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 2, pp. 272–285, Feb. 2017.

[29] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong, "Catch me if you can: Detecting pickpocket suspects from large-scale transit records," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 87–96.

[30] J. Hua, Z. Shen, and S. Zhong, "We can track you if you take the metro: Tracking metro riders using accelerometers on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 2, pp. 286–297, Feb. 2017.

[31] M. K. El Mahrsi, E. Côme, L. Oukhellou, and M. Verleysen, "Clustering smart card data for urban mobility analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 712–728, Mar. 2017.

[32] M. Yazdanpanah and M. H. Hosseinlou, "The role of personality traits through habit and intention on determining future preferences of public transport use," *Behav. Sci.*, vol. 7, no. 1, p. 8, 2017.

[33] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "Dnn-based prediction model for spatio-temporal data," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2016, pp. 92-1–92-4.

[34] F. Zhang, N. J. Yuan, Y. Wang, and X. Xie, "Reconstructing individual mobility from smart card transactions: A collaborative space alignment approach," *Knowl. Inf. Syst.*, vol. 44, no. 2, pp. 299–323, 2015.

[35] X. Kong *et al.*, "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3874–3886, May 2018.
[36] A. Imaz, K. M. N. Habib, A. Shalaby, and A. O. Idris, "Investigating the factors affecting transit user loyalty," *Public Transp.*, vol. 7, no. 1, pp. 39–60, 2015.
[37] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.

JINZHONG WANG received the B.Sc. degree in computer education from Anshan Normal University, Anshan, China, in 2002, and the M.Sc. degree in computer application technology from Liaoning University, Shenyang, China, in 2005. He is currently pursuing the Ph.D. degree with the School of Software, Dalian University of Technology, Dalian, China. Since 2005, he has been with Shenyang Sport University, Shenyang. His research interests include computational social networks, network science, data science, and mobile social networks.

XIANGJIE KONG (M'13–SM'17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor with the School of Software, Dalian University of Technology, China. He has published over 50 scientific papers in international journals and conferences (with over 30 indexed by ISI SCIE). His research interests include big traffic data, mobile computing, and cyber-physical systems. He is a Senior Member of CCF and a member of ACM. He has served as a (guest) editor for several international journals and the workshop chair or a PC member of a number of conferences.

WENHONG ZHAO (M'18) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2002. He has been with the Zhejiang University of Technology, Hangzhou, since 1991, where he is currently a Full Professor with the Ultraprecison Machining Center. His research interests include big data, embedded systems, intelligent systems, and precision machining.

AMR TOLBA received the M.Sc. and Ph.D. degrees from the Faculty of Science, Menoufia University, Egypt, in 2002 and 2006, respectively. He is currently an Associate Professor with the Faculty of Science, Menoufia University. He is currently on leave from Menoufia University to the Computer Science Department, Community College, King Saud University, Saudi Arabia. He has authored/co-authored over 30 scientific papers in international journals and conference proceedings. His main research interests include socially aware networks, Internet of Things, intelligent systems, big data, recommender systems, and cloud computing. He serves as a technical program committee member in several conferences.

ZAFER AL-MAKHADMEH received the M.Sc. and Ph.D. degrees from the Department of Computer Engineering, Faculty of Information and Computer Engineering, Kharkov National Technical University of Ukraine, in 1998 and 2001, respectively. He is currently an Assistant Professor at the Computer Science Department, Community College, King Saud University, Saudi Arabia. His main research interests include cloud computing, social network analysis, big data, and intelligent systems.

FENG XIA (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He was a Research Fellow at the Queensland University of Technology, Australia. He is currently a Full Professor with the School of Software, Dalian University of Technology, China. He has published two books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, network science, data science, and mobile social networks. He is a Senior Member of ACM and a member of AAAS. He serves as the general chair, PC chair, workshop chair, or publicity chair of a number of conferences. He is a (guest) editor of several international journals.